John von Neumann Institute for Computing

Determination of Protein Structure in Solution Based on $^{13}C^{\alpha}$ Chemical Shifts and NOE Distance Constraints

J. A. Vila, H. A. Scheraga

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 43-48, 2008.

http://www.fz-juelich.de/nic-series/volume40

Determination of Protein Structure in Solution Based on $^{13}C^{\alpha}$ Chemical Shifts and NOE Distance Constraints

Jorge A. Vila[1,2] **and Harold A. Scheraga**[1]

[1] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, U.S.A.
E-mail: has5@cornell.edu

[2] Universidad Nacional de San Luis, Instituto de Matemática Aplicada San Luis, CONICET, Ejército de Los Andes 950, 5700 San Luis, Argentina

A recently introduced physics-based method that exploits distance constraints derived from Nuclear Overhauser Effect and $^{13}C^{\alpha}$ chemical shift data aimed at determining, validating and refining, protein structures at a high level of accuracy, without resorting to other experimental data, is illustrated here by determining the native structures of two proteins, namely a 20-residue all-β and a 76-residue all-α protein. The approach makes use of $^{13}C^{\alpha}$ chemical shifts, computed at the density functional level of theory to derive the backbone and side-chain torsional constraints for *all* backbone and side-chain torsional angles *dynamically*. Consequently, this method is expected to lead to a more precise characterization of the conformational distributions for the backbone, as well as for the side chains of the amino acid residues on both the surface and in the interior of a protein. With the available computational resources, there are three main advantages of this new methodology: (a) it can be used for proteins of *any* class or size; (b) it provides a unified, self-consistent, methodology to determine, validate and refine proteins structures at a *high-quality* level; and (c) it does not use any knowledge-based information and hence, it is a purely *physics-based* method. The anticipated results of our applications indicated that, starting from randomly generated conformations, the final protein models are more accurate than existing NMR-derived models (obtained by using traditional methods) in terms of the agreement between predicted and observed $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts as well as some stereochemical quality indicators.

1 Introduction

The backbone and side-chain conformations of a residue are influenced by interactions with the rest of the protein but, once these conformations are established by these interactions, the $^{13}C^{\alpha}$ chemical shift of this residue depends, mainly, on its backbone,[1] and its side-chain[2,3] conformation, with no significant influence of either the amino acid sequence[4] or the position of the given residue in the sequence[4]. These properties, together with the fact that $^{13}C^{\alpha}$ is ubiquitous in proteins, make this nucleus an attractive candidate for computation of theoretical chemical shifts at the quantum chemical level of theory in order to determine, validate and refine protein structures[4,5]. We have been developing methodology to use $^{13}C^{\alpha}$ chemical shifts, in addition to other NMR data, to determine, validate and refine protein structure in solution[5–7].

This methodology[4], validated on 139 conformations of the human protein ubiquitin, enabled us to offer a new criterion for an accurate assessment of the quality of NMR-derived protein conformations and to examine whether X-ray or NMR-solved structures are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution. A detailed analysis[4] of the disagreement between observed and DFT-computed $^{13}C^{\alpha}$ chemical shifts in these ubiquitin conformations illustrated the accuracy of the calculations and, more

important, demonstrated that these disagreements reflect the dynamic nature of the protein, rather than inaccuracies of the method. Our methodology has also been used[6] to show that neutral, rather than charged, basic and acidic groups are a better approximation of the observed $^{13}C^{\alpha}$ chemical shifts of a protein in solution.

The goal of this work is to illustrate how this methodology (a) can be used for proteins of *any* class or size; (b) provides a unified, self-consistent, methodology to determine, validate and refine proteins structures at a *high-quality* level; and (c) does not use any knowledge-based information and hence, it is a purely *physics-based* method. To accomplish this goal, the methodology is illustrated here with two applications: first, to determine an accurate set of conformations that simultaneously satisfies the NOE-derived distance constraints and the $^{13}C^{\alpha}$-derived torsional constraints for a 20-residue peptide capable of forming a three-stranded antiparallel β-sheet in aqueous solution[8], i.e., the BS2 peptide with the sequence: TWIQN$_D$PGTKWYQN$_D$PGTKIYT, for which complete sets of both $^{13}C^{\alpha}$ chemical shifts and NOEs were reported[8]. Secondly, as an additional test of the procedure, we chose to determine the tertiary structure of the B. Subtilis Acyl Carrier (SAC) protein[5]. This is a small all-α helical protein with only 76 amino acid residues and no disulfide bonds, for which all the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts and the NOE-derived distances are available from the Biological Magnetic Resonance Data Bank under *accession number* 4989. The NMR structure of the SAC protein has been solved by Xu et al.[9] using traditional methods, and the coordinates of the average-minimized structure were deposited in the Protein Data Bank with the code 1HY8.

2 Materials and Methods

Figure (1) shows a flow chart of the protein structure determination procedure which, essentially, consists of 4 steps, namely:

(1) The Variable-Target-Function (VTF) approach with a simplified soft-sphere potential function[10] is used to generate an ensemble of conformations at random that simultaneously satisfy a set of distance constraints derived from the experimental NOEs and the backbone torsional constraints derived from the $^{13}C^{\alpha}$ conformational shifts, i.e., only for the regular α-helical and β-sheet segment of the molecule. Among all generated VTF conformations, only those possessing a maximum NOE-derived distance violation lower than a certain cut off value, e.g., 1 Å, are selected. If the number of selected conformations is greater than $\sim$10, then a clustering procedure is applied by using the Minimal Spanning Tree (MST) method[11].

(2) The $^{13}C^{\alpha}$ chemical shifts are computed at the DFT level[4–6] for each conformation of the set obtained in step (1). The DFT procedure is applied to each amino acid **X** in the sequence by treating **X** as a terminally-blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of each generated peptide structure. Examination of the chemical shifts of each residue of all the clustered conformations considered here enabled us to identify a new *minimal*-rmsd model[4] in which the $^{13}C^{\alpha}$ chemical shift of each residue individually best matched the experimental one, thereby providing a *new* set of ϕ, ψ, and χ torsional angle constraints of *all* the residues of the molecule[5].

(3) Only one conformation among all the selected conformations described in step (1), was selected. This conformation possessed the lowest rmsd between the computed and observed $^{13}C^{\alpha}$ chemical shifts. The selected conformation was used as a starting one in a

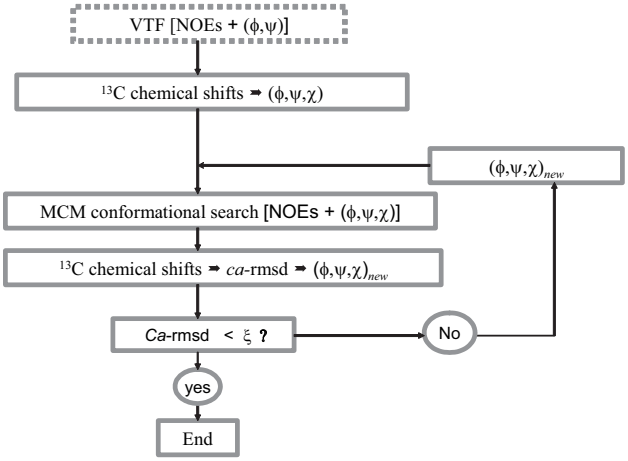


Figure 1. Flow chart illustrating the steps of the computational procedure, as described in the Materials and Methods section. VTF is the acronym for the Variable-Target-Function approach[10]. The variable ξ represents the convergence criterion.

conformational search with the Monte Carlo with Minimization (MCM)[12] method carried out with two types of constraints: the original fixed set of NOEs and the *new* set of torsional angles derived in step (2). This time, instead of using a simplified soft-sphere potential function, we use a complete force-field containing the following terms: (a) the internal potential energy, as described by the ECEPP/3 force field[13]; and (b) additional energy terms aimed at penalizing violations of the distance and torsional angle constraints[14]. Finally, a clustering procedure is carried out to select a small sub-set of the total number of the MCM-derived set of conformations by using the MST method[11] and assuming a specific rmsd cutoff for all heavy atoms.

(4) Steps (2) and (3) are repeated iteratively by using the set of conformations obtained in step (3) and, hence, enabling us to obtain an *updated* set of torsional-angle constraints. At any stage of the procedure, a tolerance range Λ, with $20^o \leq \Lambda \leq 35^o$, for the torsional constraints was adopted. Variation of the torsional angles within a tolerance range Λ is considered acceptable and hence is not subject to energetic penalties. Among all the conformations generated in the final use of step (3), only one conformation is selected, because it is characterized by the lowest rmsd between the computed $^{13}C^{\alpha}$ chemical shifts and the observed ones. Thus, the procedure of step 3, applied to such a conformation, led to a new set of structures. The final number of conformations in this set is determined by the cutoff rmsd value adopted for the clustering procedure in step (3).

Application of this procedure to 20-residue and 76-residue proteins enabled us to determine a Set_β, consisting of 10 conformations[7], and a Set_α, consisting of 9 conformations[5], respectively. Analysis of the quality of these sets of conformations in terms of the *ca*-rmsd[4] and some structural quality indicators are given in Table 1.

Conformation Set[a]		*Ca*-rmsd[b] [ppm]	Maximum Distance Violation[c] [Å]	Structural Quality Indicators[d]
all-β	Santiveri (20)[8]	4.6	2.36	61 ± 11 (39 ± 12) [0.0]
	Set_β (10)[7]	3.5	0.88	62 ± 10 (37 ± 9) [0.0]
all-α	1HY8 (1)[9]	3.9	1.38	95.8 (2.8) [1.4]
	Set_α (9)[5]	2.9	0.63	92.0 ± 1 (4.3 ± 0.9) [1.5]

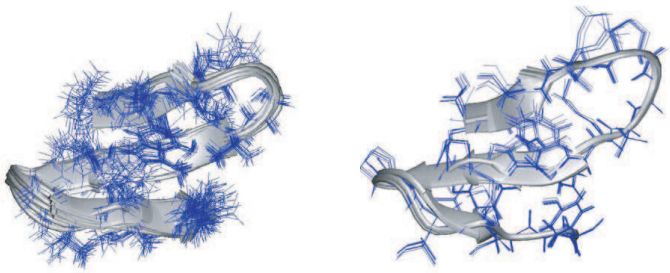
Table 1. Results for the all-α and all-β protein structure determination.

[a] Computed for each set of conformations listed; the number of conformations in each set is indicated in parentheses. Set_β[7] and Set_α[5] refer to the set of conformations determined as explained in the Materials and Method section.
[b] Values computed as explained in Vila et al.[4]
[c] From the *full* set of NOE-derived distances, namely 130 for the 20-residue all-β and 1,050 for the 76-residue all-α proteins, respectively.
[d] Based on PROCHECK[15]. The listed values are the number of residues in the allowed regions of the Ramachandran map; in the generously allowed regions (in parenthesis); and in the disallowed regions (in bracket). All the listed values, except for the protein 1HY8, are averaged over the total number of conformations of each set.

3 Results and Discussion

The results obtained here indicate that an *all* β-sheet (Figure 2a-b) and an *all* α-helical (Figure 3) set of structures can be determined by simply identifying a set of conformations which simultaneously satisfy a set of constraints, namely $^{13}C^{\alpha}$-dynamically-derived torsional angle constraints for *all* amino acid residues in the sequence, and a fixed set of NOE-derived distance constraints. Analysis of the accuracy of these sets, as a measure of the closeness with which the calculations reproduce the structure in solution, in terms of the NOE-derived maximum distance violations, the $^{13}C^{\alpha}$ chemical shifts, and some stereochemical quality factors (see Table 1), indicates that our self-consistent physics-based method is able to produce a more accurate set of conformations than that obtained with the traditional methods.

In summary, these applications illustrate the three main advantages of this new methodology: (a) it can be used for proteins of *any* class or size; (b) it provides a unified, self-consistent, methodology to determine, validate and refine protein structures at a *high-quality* level; and (c) it does not use any knowledge-based information and, hence, it is a purely *physics-based* method.



(a) **(b)**

Figure 2. (**a**) Superposition of 20 NMR-derived conformations (represented by ribbon diagrams) of the BS2 peptide obtained by Santiveri *et al*[8]. Side chains are represented by thin black lines. (**b**) Same as (**a**) for the 10 NMR-derived conformations in this work from Set_β[7].

Figure 3. Ribbon diagram of the superposition of nine models of the Set_α for the SAC protein[5] (in grey color) and the minimized average NMR structure (1HY8) [in black][9].

Acknowledgments

This research was supported by grants from the National Institutes of Health (GM-14312 and GM-24893), and the National Science Foundation (MCB05-41633). Support was also received from the CONICET, FONCyT-ANPCyT (PAV 22642 / 22672), and from the Universidad Nacional de San Luis (P-328501), Argentina. The research was conducted using the resources of a Beowulf-type cluster located at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University; and the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center.

References

1. S. Spera and A. Bax Empirical correlation between protein backbone conformation and C^{α} and C^{β} ^{13}C Nuclear Magnetic Resonance chemical shifts. J. Am. Chem Soc., 113: 5490-5492, 1991.
2. R.H Havlin, H. Le, D.D. Laws, A.C. deDios and E. Oldfield. An ab initio quantum chemical investigation of carbon-13 NMR shielding tensors in glycine, alanine,

valine, isoleucine, serine, and threonine: Comparisons between helical and sheet tensors, and effects of χ_1 on shielding. J Am Chem Soc., 119: 11951-11958, 1997.
3. M.E. Villegas, J.A. Vila and H.A. Scheraga. Effects of Side-Chain Orientation on the ^{13}C Chemical Shifts of Antiparallel β-sheet Model Peptides. J Biomol NMR, 37: 137-146, 2007.
4. J.A. Vila, M.E. Villegas, H.A. Baldoni and H.A. Scheraga. Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures. J. Biomol. NMR, 38:221-235, 2007
5. J.A. Vila, D.R. Ripoll and H.A. Scheraga. Use of $^{13}C^{\alpha}$ chemical shifts in protein structure determination. J. Phys. Chem. B, 111:6577-6585, 2007.
6. J.A. Vila and H.A. Scheraga. Factors affecting the use of $^{13}C^{\alpha}$ chemical shifts to determine, refine, and validate protein structures. Proteins 71**,** 641-654, 2008.
7. J.A. Vila, Y.A. Arnautova, and H.A. Scheraga. Use of $^{13}C^{\alpha}$ chemical shifts for accurate determination of β-Sheet structures in solution. Proc. Natl. Acad. Sci. USA**,** 105, 1891-1896, 2008.
8. C.M. Santiveri, J. Santoro, M. Rico and M.A. Jiménez. Factors involved in the stability of isolated beta-sheets: turn sequence, beta-sheet twisting, and hydrophobic surface burial. Prot. Sci., 13:1134-1147, 2004.
9. G.-Y. Xu, A. Tam, L. Lin, J. Hixon, C.C. Fritz and R. Powers. Solution structure of B. Subtilis Acyl carrier protein. Structure, 9:277-287, 2001.
10. M. Vásquez and H.A. Scheraga. Variable-Target-Function and buildup procedures for the calculation of protein conformation – application to bovine pancreatic trypsin-inhibitor using limited simulated Nuclear Magnetic-Resonance data**.** J. Biomol. Struct. Dyn., 5:757-784, 1998.
11. J.B. Kruskal, Jr., On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proc. American. Math. Soc., 7:48-50, 1956.
12. Z. Li and H.A. Scheraga. Monte-Carlo-Minimization approach to the multiple-minima problem in protein folding. Proc. Natl. Acad. Sci. USA, 84:66116615, 1987.
13. G. Némethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H.A. Scheraga. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem., 96:6472-6484, 1992.
14. D.R. Ripoll and F. Ni. Refinement of the thrombin-bound structure of a hirudin peptide by a restrained Electrostatically Driven Monte-Carlo Method. Biopolymers, 32:359-365, 1992.
15. R.A. Laskowski, M.W. MacArthur, D.S. Moss and J. Thornton. PROCHECK - a program to check the stereochemical quality of protein structures. J. Appl. Cryst.,26:283-291, 1993.